Stein's Method

Matthew Kwan

March 9, 2014

Contents

1 Introduction

Comment 1.1. The plan is to introduce with limit theorems: Central Limit theorem, Poisson Limit theorem. The failure of limit theorems is that they provide no understanding of speed of convergence, in particular convergence cannot be assumed to be uniform as parameters vary.

Stein's method is a technique for bounding the distance between distributions, with a variety of different distance metrics. Quantitative bounds can be useful in their own right, or can be further applied to prove asymptotic results.

1.1 Notation

For this thesis, the set of natural numbers $\mathbb{N}$ includes zero. We write $\mathbf{1}_A$ for the characteristic function of a set A: $\mathbf{1}_A(x) = 1$ if $x \in A$, otherwise $\mathbf{1}_A(x) = 0$. Also, $[k]$ denotes the set $\{1, \ldots, k\}$.

Unless otherwise specified, all asympotics are as $n \to \infty$. Apart from standard asymptotic notation, we use two notions of asymptotic equivalence: $f \sim g$ means $f = g(1 + o(1))$ and $f \asymp g$ means $f = \Theta(g)$.

In this thesis, unless stated otherwise, graphs are labelled. That is, they are distinguished even within isomorphism classes. A graph may not have loops or multiple edges; an object which is allowed to have loops and/or multiple edges will be called a multigraph.

The phrase "randomly choose" is taken to mean a uniformly random choice. That means, each possible option is chosen with equal probability.

Part I

Theory

2 General Probability Theory

2.1 Probability Spaces

Comment 2.1. I'm a little bit uncertain how much depth to go into for this. At the moment, it's written so that someone who's seen measure theory but no probability theory (an analyst) can understand. Where possible, I've tried to translate things into the discrete case, because it's often more intuitive (and since I plan for applications to be combinatorial).

For many combinatorial applications, an informal understanding of probability theory will suffice. However, in this thesis a rigorous foundation in probability theory will be useful. The following is intended only as a brief review.

Definition 2.1. A *probability space* is a measure space $(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathbb{P}(\Omega) = 1$. In this case we say $\mathbb{P}$ is a *probability measure*, and denote the set of all probability measures on $(\Omega, \mathcal{A})$ by $\mathcal{P}(\Omega, \mathcal{A})$ or $\mathcal{P}(\Omega)$ if there is no ambiguity. An *event* is a measurable set $A \in \mathcal{A}$.

For our purposes Ω will often be countable, with $\mathcal{A}$ as the power set of Ω. In this case $\mathbb{P}$ is uniquely defined by the probabilities $\mathbb{P}(\omega) := \mathbb{P}(\{\omega\})$, for each $\omega \in \Omega$. We will discuss specific probability spaces on combinatorial objects in Section 3.

For an event A, $\mathbb{P}(A)$ is interpreted as the "probability that A occurs". Events will usually be of the form $A = \{\omega \in \Omega : P(\omega) \text{ holds}\}$, where $P(\omega)$ is some property of an object ω. For clarity, we often abuse notation slightly and write $\mathbb{P}(P \text{ holds})$ instead of $\mathbb{P}(\{\omega \in \Omega : P(\omega) \text{ holds}\})$.

2.2 Random Elements

Definition 2.2. A *random element* is a measurable function $X : (\Omega_1, \mathcal{A}_1) \to (\Omega_2, \mathcal{A}_2)$ between measurable spaces. If $\Omega_2 = \mathbb{R}^n$ for some $n \in \mathbb{N}$, with $\mathcal{A}_2$ the Borel σ-algebra on $\mathbb{R}^n$, then we say X is a *random vector*. A one-dimensional random vector is a *random variable*. If Ω_2 is countable then we say X is *discrete*.

Especially in combinatorial spaces, Ω_1 is often countable. In this case, any function from a probability space $\big(\Omega_1, 2^{\Omega_1}, \mathbb{P}\big)$ is measurable.

To interpret a random variable, we need a probability measure $\mathbb{P}$ on the underlying measurable space $(\Omega_1, \mathcal{A}_1)$ (often, this will be implicit). Then, $\mathbb{P}(X \in A) = \mathbb{P}(\{\omega \in \Omega_1 : X(\omega) \in A\})$ is the probability that X takes a value in the set A. Often, we will only be interested in such probabilities: that is, we do not care about the realization of a random variable as a function on an underlying probability space. This motivates the following definition:

Definition 2.3. Suppose X is a random element which takes values in the measurable space $(\Omega, \mathcal{A})$. The *distribution* (or *law*) $\mathcal{L}_X$ of X with respect to an underlying probability $\mathbb{P}$ is the pushforward measure with respect to X. That is, it is a probability measure defined by $\mathcal{L}_X(A) = \mathbb{P}\big(X^{-1}(A)\big)$ for $A \subseteq \mathcal{A}$. Also, we occasionally use the notation $\mathcal{L}(X) := \mathcal{L}_X$ for ease of reading.

It is worth noting that in fact any probability measure is the distribution of some random element, so we can define a probability distribution in the abstract and then assert the existence of a random variable with that distribution. To see this, note that given a probability measure $\mathcal{L} \in \mathcal{P}(\Omega)$, we can choose $X = \mathrm{id}_\Omega$ to have $\mathcal{L}_X = \mathcal{L}$ with respect to the underlying probability measure $\mathcal{L}$. We also use the notation $X \in \mathcal{L}$ to indicate that X has distribution $\mathcal{L}$.

Example 2.4. The normal distribution with parameters μ and σ is denoted $\mathcal{N}(\mu, \sigma)$ or $\mathcal{N}_{\mu,\sigma}$ and is defined by $\mathcal{N}_{\mu,\sigma}(B) = \frac{1}{\sigma\sqrt{2\pi}} \int_B e^{-\frac{(x-\mu)^2}{2\sigma^2}} \,\mathrm{d}\, x$ for any Borel set B.

The Poisson distribution with parameter λ is denoted $\mathrm{Po}(\lambda) = \mathrm{Po}_\lambda$; this is defined by $\mathrm{Po}_\lambda(k) = \frac{\lambda^k e^{-\lambda}}{k!}$ for all $k \in \mathbb{N}$.

It is in general a little tricky to define "the set of values a random element can take", but this is straightforward in the discrete case.

Definition 2.5. The *support* of a discrete random element X is the set

$$\text{supp}(X) = \{k \in \Omega : \mathbb{P}(X = k) > 0\}.$$

For example, the Poisson distribution has support $\mathbb{N}$.

2.3 Dependence and Coupling

Definition 2.6. Suppose X and X' are random elements $(\Omega_1, \mathcal{A}_1, \mathbb{P}) \to (\Omega_2, \mathcal{A}_2)$. We say that X and X' are *independent* if $\mathbb{P}(X \in A_2)\mathbb{P}(X' \in A_2) = \mathbb{P}(X \in A_2 \text{ and } X' \in A_2)$ for all $A_2 \in \mathcal{A}_2$.

If $A_1, A_1' \in \mathcal{A}$ then we say A_1 and A_1' are independent if $\mathbb{P}(A_1)\mathbb{P}(A_1') = \mathbb{P}(A_1 \cap A_1')$. We can similarly say an event is independent of a random element.

If two objects are not independent, then we say they are *dependent*.

Intuitively, two objects are dependent if information about one object can give information about the other. For example, we might be interested in the probability of an event A, under the assumption that the event A' occurs.

Definition 2.7. The *conditional probability* of an even A given an event A' with nonzero probability is $\mathbb{P}(A|A') = \mathbb{P}(A \cap A')/\mathbb{P}(A')$

We can also condition random elements on an event.

Definition 2.8. Suppose $X : (\Omega_1, \mathcal{A}_1, \mathbb{P}) \to (\Omega_2, \mathcal{A}_2)$ is a random element, and $A_1 \in \mathcal{A}_1$ is an event with nonzero probability. Then the *distribution of X conditioned on A_1* is denoted by $\mathcal{L}_{X|A_1}$ and defined by $\mathcal{L}_{X|A_1}(A_2) = \mathbb{P}(X \in A_2|A_1)$ for $A_2 \in \mathcal{A}_2$.

Given a finite collection of measure spaces $(\Omega_1, \mathcal{A}_1, \mu_1), \ldots, (\Omega_n, \mathcal{A}_n, \mu_n)$, recall the construction of the product measure space $(\Omega, \mathcal{A}, \mu) := (\prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \mathcal{A}_i, \prod_{i=1}^n \mu_i)$. If a random element takes values in a product space then each component is measurable, and conversely if the components of a random tuple are measurable then that tuple is measurable in the product space. So, we can make the following definitions:

Definition 2.9. Given random elements $X_1, \ldots, X_n$ on the same underlying probability space, $\mathcal{L}(X_1, \ldots, X_n) := \mathcal{L}((X_1, \ldots, X_n))$ is called the *joint distribution* of $X_1, \ldots, X_n$. Conversely, given a random tuple $(X_1, \ldots, X_n)$, each $\mathcal{L}(X_i)$ is called a *marginal distribution*.

Suppose we have two distributions of random elements $\mathcal{L}(X_1)$ and $\mathcal{L}(X_2)$. *Coupling* is the technique of constructing a random ordered pair (X_1, X_2) which realizes the given distributions as marginal distributions. Usually this is done by specifying the joint distribution $\mathcal{L}(X_1, X_2)$.

The idea is that coupling creates a particular kind of dependence between X_1 and X_2 that allows us to compare the two distributions. Often, we are able to make conclusions about the distributions $\mathcal{L}(X_i)$ which are independent of their specific realizations as random elements in the coupling.

2.4 Expected Value

Definition 2.10. The *expected value* of a random variable X is $\mathbb{E}X = \int x \, \mathrm{d}\, \mathcal{L}_X(x)$.

For a random variable X that takes integer values, this definition is equivalent to the well-known formula $\mathbb{E}X = \sum_{x \in \mathbb{Z}} x \, \mathbb{P}(X = x)$.

Remark 2.11. If X is a random variable that can be interpreted as counting the number of objects that satisfy some property, then we can express X as a sum of indicator variables $\sum_i \mathbf{1}_{A_i}$, where A_i is the event that the ith object satisfies our property. Noting that $\mathbb{E}$ is linear, we have $\mathbb{E}X = \sum_i \mathbb{E}\,\mathbf{1}_{A_i} = \sum_i \mathbb{P}A_i$. So, in order to compute the expectation of X we just need to compute the probability that each object satisfies our required property.

If we fix a particular underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$, we can also equivalently view expectation as a linear functional on the space of integrable functions: $\mathbb{E}f = \int f(\omega)\,\mathrm{d}\,\mathbb{P}$. Sometimes we will define a new probability space $(\Omega, \mathcal{A}, \mathbb{P}')$ on an existing measurable space. In this case we will write $\mathbb{E}_{\mathbb{P}'}$ to indicate expecation with respect to the measure $\mathbb{P}'$, to avoid ambiguity.

In fact, probability measures are uniquely determined by their expectation functional, because $\mathbb{E}_{\mathbb{P}}\,\mathbf{1}_A = \mathbb{P}(A)$ for all events A.

Definition 2.12. A set of real functions $\mathcal{H}$ is a *determining class* if $\mathbb{E}_{\mathbb{P}_1} h = \mathbb{E}_{\mathbb{P}_2} h$ for all $h \in \mathcal{H}$ implies that $\mathbb{P}_1 = \mathbb{P}_2$.

We can also define the expectation functional of a random variable $\mathbb{E}_X := \mathbb{E}_{\mathcal{L}(X)}$.

Definition 2.13. The expected value of a random variable with distribution $\mathcal{L}_{X|A_1}$ is called the *conditional expected value of X given A_1* and is denoted $\mathbb{E}[X|A_1]$.

We can also define conditional expectation with respect to another random variable. If X_1 and X_2 are random variables defined on the same underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$, then the sets $X_2^{-1}(B)$ for Borel B comprise a sub-σ-algebra $\mathcal{A}'$ of $\mathcal{A}$. Then, $\mu : A' \mapsto \mathbb{E}[X_1\,\mathbf{1}_{A'}]$ is a signed measure on $\mathcal{A}'$ that is absolutely continuous with respect to the restriction of $\mathbb{P}$ to $\mathcal{A}'$. By the Radon-Nikodym theorem there is an $\mathcal{A}'$-measurable random variable $\mathbb{E}[X_1|X_2]$ that satisfies $\mathbb{E}[X_1\,\mathbf{1}_{A'}] = \mathbb{E}[\mathbb{E}[X_1|X_2]\,\mathbf{1}_{A'}]$ for all A' in $\mathcal{A}'$. This random variable is almost uniquely defined: for any two choices of $\mathbb{E}[X_1|X_2]$, the probability that they differ is zero.

Definition 2.14. The random variable $\mathbb{E}[X_1|X_2]$ as defined above is called the *conditional expectation of X_1 with respect to X_2*. We can also view conditional expectation as a linear operator between functions: we define $\mathbb{E}^{X_2}$ by $X_1 \mapsto \mathbb{E}[X_1|X_2]$.

This definition generalizes the previous definition of expectation conditioned on an event: if $\omega \in A$ and $\mathbb{P}(A) > 0$ then $\mathbb{E}[X|\,\mathbf{1}_A](\omega) = \mathbb{E}[X|A]$.

Note that if X_2 is discrete then we do not need to invoke Radon-Nikodym. We can define $\mathbb{E}[X_1|X_2]$ by $\mathbb{E}[X_1|X_2](\omega) = \mathbb{E}[X_1|X_2 = X_2(\omega)]$ for all $\omega \in \Omega$ with $\mathbb{P}(X_2 = X_2(\omega)) > 0$; this defines $\mathbb{E}[X_1|X_2]$ up to a set of probability zero.

We finally present a simple consequence of the definition of conditional expectation.

Proposition 2.15 (Tower Law of Expectation)**.** *Suppose X_1 and X_2 are random variables defined on the same underlying probability space $(\Omega, \mathcal{A})$. Then $\mathbb{E}\left[\mathbb{E}^{X_2} X_1\right] = \mathbb{E}[X_1]$.*

Proof. $\mathbb{E}\left[\mathbb{E}^{X_2} X_1\right] = \mathbb{E}[\mathbb{E}[X_1|X_2]\,\mathbf{1}_\Omega] = \mathbb{E}[X_1\,\mathbf{1}_\Omega] = \mathbb{E}[X_1]$ □

2.5 Markov Chains

Comment 2.2. I'll need to define Markov Chains, stationary distributions, irreducibility and time-reversibility.

Perhaps I should talk more generally about stochastic processes, because applying exchangeable pairs to Stein's method is has connections with Ornstein-Uhlenbeck processes and also Stein's method can be applied to Poisson processes.

2.6 The Weak Topology on Probability Measures

Comment 2.3. The main purpose of this section is to motivate the metrics usually used in Stein's method: they are all legitimate topological metrics and are consistent with the topology of convergence in distribution. In particular, if we can show $\mathrm{d}_{\mathcal{H}}(X_n, X) \to 0$ we have shown that $X_n \xrightarrow{d} X$, as Toby does [Joh11].

Definition 2.16. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. We say X_n *converges in distribution* to a random variable X if $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for all bounded continuous functions f. Alternatively, we say $\mathcal{L}(X_n)$ converges *weakly* to $\mathcal{L}(X)$, or simply $\mathcal{L}(X_n) \to \mathcal{L}(X)$. The topology on $\mathcal{P}(\mathbb{R})$ associated with this convergence is called the *weak topology* (we will see that it is indeed a topology). Convergence in distribution of random vectors is defined componentwise.

Definition 2.17. The *distribution function* F_X of a random variable X is defined by $F_X(x) = \mathbb{P}(X \leq x)$.

Theorem 2.18. *The following are equivalent.*

(i) $\mathcal{L}(X_n) \to \mathcal{L}(X)$

(ii) $F_{X_n}(x) \to F_X(x)$ *for all* x *where* F_X *is continuous*

(iii) *(Lévy's continuity theorem)* $\mathbb{E}e^{itX_n} \to \mathbb{E}e^{itX}$ *for all* $t \in \mathbb{R}$.

The equivalence of Conditions (i) and (ii) is a well-known result called the Portmanteau Theorem.

When X and each X_n are integer random variables, then Condition (ii) reduces to the condition that $\mathbb{P}(X_n = k) \to \mathbb{P}(X = k)$ for all k. This characterization is usually used to prove the Poisson limit theorem.

Classically, distributional convergence results are often proved by Lévy's continuity theorem. For example, this approach is usually used to prove the central limit theorem. For combinatorial applications, convergence in distribution can also be proved by the "method of moments": if X is the only random variable with the moments $\left(\mathbb{E}X^k\right)_{k\in\mathbb{N}}$, then $\mathcal{L}(X_n) \to \mathcal{L}(X)$ if $\mathbb{E}X_n^k \to \mathbb{E}X^k$. Convergence in distribution can also sometimes be inferred from stronger forms of convergence when X and all the X_n are coupled to the same underlying space.

A disadvantage of all these approaches is that they provide little information about the rate of convergence.

In functional analysis terms, note that expectation operators are bounded linear functionals on the space of real bounded continuous functions. Then, $\mathcal{L}(X_n) \to \mathcal{L}(X)$ just means that $\mathbb{E}_{X_n} \to \mathbb{E}_X$ in the weak-star topology. Although $C_b(\mathbb{R})^*$ is not metrizable, the subspace corresponding to $\mathcal{P}(\mathbb{R})$ is in fact metrizable, with a metric called the Lévy metric. For Stein's method we will be interested in some slightly stronger metrics, which we will now define.

Definition 2.19. Let $\mathcal{H}$ be a collection of real measurable "test" functions, and let $S \subseteq \mathcal{P}(\mathbb{R})$ be a set of "comparable" measures. Define $d_{\mathcal{H}} : S^2 \to \mathbb{R}^+$ by $d_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_2) = \sup_{h\in\mathcal{H}} |\mathbb{E}_{\mathbb{P}_1} h - \mathbb{E}_{\mathbb{P}_2} h|$. For random variables X_1, X_2, we write $d_{\mathcal{H}}(X_1, X_2)$ instead of $d_{\mathcal{H}}(\mathcal{L}(X_1), \mathcal{L}(X_2))$.

Remark 2.20. In order for Definition 2.19 to be well-defined, we should choose $\mathcal{H}$ and S such that $\mathbb{E}_{\mathbb{P}}h$ is well-defined and finite for all $\mathbb{P} \in S$ and $h \in \mathcal{H}$. For $d_{\mathcal{H}}$ to be a metric, we should also ensure that $d_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_2) < \infty$ for all $\mathbb{P}_1, \mathbb{P}_2 \in S$.

Modulo the considerations in Remark 2.20, $d_{\mathcal{H}}$ is always non-negative, symmetric and satisfies the triangle inequality. Hence, to check that $d_{\mathcal{H}}$ is a metric, we only need to check that $d_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_2) = 0$ implies that $\mathbb{P}_1 = \mathbb{P}_2$. It suffices to check that $\mathcal{H}$ is a determining class (Definition 2.12).

Definition 2.21. We define some special cases of $d_{\mathcal{H}}$.

- If $\mathcal{H}_{\mathrm{K}} = \left\{\mathbf{1}_{(-\infty,x]} : x \in \mathbb{R}\right\}$ then $d_{\mathrm{K}} := d_{\mathcal{H}_{\mathrm{K}}}$ is called the *Kolmogorov metric*. We can choose $S_{\mathrm{K}} = \mathcal{P}(\mathbb{R})$; that is, all probability measures are comparable.
- If $\mathcal{H}_{\mathrm{W}}$ is the set of real functions h that satisfy $|h(x_1) - h(x_2)| \leq |x_1 - x_2|$ for all $x_1, x_2 \in \mathbb{R}$ (that is, the set of functions with Lipschitz constant 1), then $d_{\mathrm{W}} := d_{\mathcal{H}_{\mathrm{W}}}$ is called the *Wasserstein metric*. We can only compare absolutely integrable probability measures. That is, let $S_{\mathrm{W}} \subset \mathcal{P}(\mathbb{R})$ be the set of probability measures $\mathbb{P}$ that satisfy $\mathbb{E}_{\mathbb{P}}|\mathrm{id}_{\mathbb{R}}| < \infty$.
- If $\mathcal{H}_{\mathrm{TV}}$ is the set of functions $\mathbf{1}_B$ for Borel B, then $d_{\mathrm{TV}} := d_{\mathcal{H}_{\mathrm{TV}}}$ is called the *total variation metric*. We can choose $S_{\mathrm{TV}} = \mathcal{P}(\mathbb{R})$.

Proposition 2.22. *The Kolmogorov, Wasserstein and total variation "metrics" are actually metrics.*

Proof. First we consider the case $\mathcal{H} \in \{\mathcal{H}_{\mathrm{K}}, \mathcal{H}_{\mathrm{TV}}\}$. Since the functions in $\mathcal{H}$ are uniformly absolutely bounded (by 1), the conditions in Remark 2.20 are immediately satisfied. Now, note that the sets $(-\infty, x]$ for $x \in \mathbb{R}$ generate the Borel σ-algebra. Since $\mathbb{P}((-\infty, x]) = \mathbb{E}_{\mathbb{P}}\,\mathbf{1}_{(-\infty,x]}$ and $\mathbf{1}_{(-\infty,x]} \in \mathcal{H}$ for all $x \in \mathbb{R}$ and $\mathbb{P} \in \mathcal{P}(\mathbb{R})$, it follows that $\mathcal{H}$ is a determining class. We conclude that d_{K} and d_{TV} are metrics.

Now, $\mathbb{E}_{\mathbb{P}}|h| \leq |h(0)| + \mathbb{E}_{\mathbb{P}}|\mathrm{id}_{\mathbb{R}}| < \infty$ for $\mathbb{P} \in S_{\mathrm{W}}$ and $h \in \mathcal{H}_{\mathrm{W}}$. So, $\mathbb{E}_{\mathbb{P}}h$ is well-defined and finite. If $\mathbb{P}_1, \mathbb{P}_2 \in S_{\mathrm{W}}$ then $d_{\mathrm{W}}(\mathbb{P}_1, \mathbb{P}_2) \leq 2\mathbb{E}_{\mathbb{P}}|\mathrm{id}_{\mathbb{R}}| < \infty$. So, the conditions in Remark 2.20 are satisfied for the Wasserstein "metric".

Next, let $\mathbb{P}_1, \mathbb{P}_2 \in S_{\mathrm{W}}$ and suppose $\mathbb{E}_{\mathbb{P}_1} h = \mathbb{E}_{\mathbb{P}_1} h$ for all $h \in \mathcal{H}_{\mathrm{W}}$. For $\varepsilon > 0$ and $x \in \mathbb{R}$, let $h_{x,\varepsilon}$ be the continuous function which takes the value 1 on the set $(-\infty, x]$, takes the value 0 on the set $[x+\varepsilon, \infty)$, and is linearly interpolated in the range $[x, x+\varepsilon]$. Since $\varepsilon h_{x,\varepsilon} \in \mathcal{H}_{\mathrm{W}}$, we have $\mathbb{E}_{\mathbb{P}_1} h_{x,\varepsilon} = \mathbb{E}_{\mathbb{P}_2} h_{x,\varepsilon}$ for each $\varepsilon > 0$. For each $x \in \mathbb{R}$, $h_{x,1/n} \to \mathbf{1}_{(-\infty,x]}$ pointwise and each $h_{x,1/n} \leq 1$ so by the dominated convergence theorem, $\mathbb{E}_{\mathbb{P}_i} h_{1/n} \to \mathbb{E}_{\mathbb{P}_i} \mathbf{1}_{(-\infty,x]}$ for each $i \in \{1, 2\}$. We have proved that $\mathbb{P}_1((-\infty, x]) = \mathbb{P}_2((-\infty, x])$ for all $x \in \mathbb{R}$, so $\mathbb{P}_1 = \mathbb{P}_2$ and $\mathcal{H}_{\mathrm{W}}$ is a determining class. We conclude that d_{W} is a metric. □

Proposition 2.23. *The topologies induced by the Kolmogorov, Wasserstein and total variation metrics are each stronger than the weak topology.*

Proof. We show that $d_{\mathcal{H}}(X_n, X) \to 0$ implies $\mathcal{L}(X_n) \to \mathcal{L}(X)$ for $\mathcal{H} \in \{\mathcal{H}_{\mathrm{K}}, \mathcal{H}_{\mathrm{W}}, \mathcal{H}_{\mathrm{TV}}\}$.

If $d_{\mathrm{K}}(X_n, X) \to 0$ or $d_{\mathrm{TV}}(X_n, X) \to 0$ then $F_{X_n} \to F_X$ uniformly, so certainly Condition (ii) of Theorem 2.18 holds.

Now, suppose $d_{\mathrm{W}}(X_n, X) \to 0$. Let $d_n = \sqrt{d_{\mathrm{W}}(X_n, X)}$ and recall the definition of $h_{x,\varepsilon}$ from the proof of Proposition 2.22. Since $d_n h_{x,d_n} \in \mathcal{H}_{\mathrm{W}}$ for each $n \in \mathbb{N}$, we have

$$\mathbb{E}_{X_n} h_{x,d_n} - \mathbb{E}_X h_{x,d_n} \leq d_{\mathrm{W}}(X_n, X)/d_n = d_n \to 0$$

uniformly for $x \in \mathbb{R}$. Now, note that

$$F_X(x-\varepsilon) \leq \mathbb{E}_X h_{x-\varepsilon,\varepsilon} \leq F_X(x) \leq \mathbb{E}_X h_{x,\varepsilon} \leq F_X(x+\varepsilon)$$

for any random variable X. If F_X is continuous at x then

$$\begin{aligned} F_{X_n}(x) - F_X(x) &\leq (\mathbb{E}_{X_n} h_{x,d_n} - \mathbb{E}_X h_{x,d_n}) + (F_X(x+d_n) - F_X(x)) \to 0 \\ F_{X_n}(x) - F_X(x) &\geq (\mathbb{E}_{X_n} h_{x-d_n,d_n} - \mathbb{E}_X h_{x-d_n,d_n}) + (F_X(x-d_n) - F_X(x)) \to 0 \end{aligned}$$

so Condition (ii) of Theorem 2.18 holds. □

Proposition 2.23 tells us that we can sensibly use our metrics to quantify the distance between random variables, in a way that is consistent with distributional (weak) convergence. All three

metrics are meaningful in their own right, but sometimes one may be easier to work with. It is sometimes possible to transfer results between metrics, though this usually results in worse constants than working directly in the desired metric. We now prove some transfer results.

Issue 2.4. It may be worthwhile to actually characterize the Wasserstein, Kolmogorov and Total Variation topologies. In particular, Wikipedia says that Wasserstein convergence is just weak convergence plus convergence of the first moment.

Also, http://en.wikipedia.org/wiki/Talk:Convergence_of_measures says an additional uniform boundedness condition on the test functions metrizes the weak topology. If so why arent we using that?

Definition 2.24. If $F_X(x) = \int_{-\infty}^{x} f_X(x) \,\mathrm{d}\, x$ for some f_X, then f_X is called the *Lebesgue density* of X, and X is called a *continuous* random variable.

If X is a continuous random variable, then by the Radon-Nikodym chain rule $\mathbb{E}_X h = \int_{\mathbb{R}} h(x) f_X(x) \,\mathrm{d}\, x$.

Proposition 2.25. *Let X_1, X_2 be random variables.*

(i) $d_{\mathrm{K}}(X_1, X_2) \leq d_{\mathrm{TV}}(X_1, X_2)$

(ii) *If $|f_{X_2}(x)| \leq C$ for all x, then $d_{\mathrm{K}}(X_1, X_2) \leq \sqrt{2C d_{\mathrm{W}}(X_1, X_2)}$.*

Proof. (Adapted from [Ros11, Proposition 1.2]). Item (i) is immediate from the definition. Then, as in the proof of Proposition 2.23,

$$\begin{aligned} F_{X_n}(x) - F_X(x) &\leq (\mathbb{E}_{X_n} h_{x,\varepsilon} - \mathbb{E}_X h_{x,\varepsilon}) + (\mathbb{E}_X h_{x,\varepsilon} - F_X(x)) \\ &\leq d_{\mathrm{W}}(X_1, X_2)/\varepsilon + \int_x^{x+\varepsilon} h_{x,\varepsilon} \, f_X(x) \,\mathrm{d}\, x \\ &\leq d_{\mathrm{W}}(X_1, X_2)/\varepsilon + C\varepsilon/2 \end{aligned}$$

and similarly

$$F_{X_n}(x) - F_X(x) \geq -d_{\mathrm{W}}(X_1, X_2)/\varepsilon - C\varepsilon/2,$$

So, we can take $\varepsilon = \sqrt{2 d_{\mathrm{W}}(X_1, X_2)/C}$ to prove Item (ii). □

Example 2.26. If $\mathcal{L}_{X_2} = \mathcal{N}(0,1)$ then $f_{X_2}(x) = (2\pi)^{-1/2} e^{-x^2/2}$ so we can take $C = (2\pi)^{-1/2}$ to obtain $d_{\mathrm{K}} \leq (2/\pi)^{1/4} \sqrt{d_{\mathrm{W}}(X_1, X_2)}$.

3 Random Combinatorial Structures

Definition 3.1. Given a finite space of combinatorial objects Ω, a probability space $\left(\Omega, 2^{\Omega}, \mathbb{P}\right)$ is often called a *model* of Ω.

Definition 3.2. In a probability space $\left(\Omega, 2^{\Omega}, \mathbb{P}\right)$ where Ω is finite, if $\mathbb{P}(\omega) = 1/|\Omega|$ for each $\omega \in \Omega$, then we say the space is *uniform*.

Uniform models are the simplest examples of random structures. For example, the uniform space $\mathcal{S}_n$ of permutations on n elements has $\mathbb{P}(\sigma) = 1/n!$ for each $\sigma \in S_n$. The uniform random graph model $\mathcal{G}_{n,M}$ has $\mathbb{P}(G) = \binom{\binom{n}{2}}{M}^{-1}$ for each graph G on the vertex set $[n]$ which has M edges. The uniform random regular graph model $\mathcal{G}_{n,d}$ is uniform on the set of all d-regular graphs on the vertex set $[n]$, though an explicit formula for the number of such graphs is not known.

As an important example of a (generally) non-uniform model, the (Erdős-Rényi) binomial random graph model $\mathcal{G}_{n,p}$ has

$$\mathbb{P}(G) = p^{|E(G)|}(1-p)^{\binom{n}{2}-|E(G)|}$$

for each graph G on the vertex set $[n]$. When $p = 1/2$, we obtain the uniform model on all graphs on the vertex set $[n]$.

One way to conceptualize the binomial model is to consider a sequence of independent coin tosses, where the coin is biased to land heads with probability p. Each coin toss corresponds to a particular potential edge, and determines whether that edge is present in the final random graph. When we define more complicated random models, we will often use this kind of informal description rather than giving an explicit formula for each $\mathbb{P}(\omega)$.

As another example, the uniform model $\mathcal{G}_{n,M}$ can be alternatively defined recursively: $\mathcal{G}_{n,0}$ is always the trivial graph with no edges, and for each $M > 0$, to obtain $\mathcal{G}_{n,M}$ we choose $G \in \mathcal{G}_{n,M-1}$ and add one of the $\binom{n}{2} - (M-1)$ possible edges at random.

Comment 3.1. This section is unfinished, I'll probably want random matrices and maybe the pairing model on random regular graphs

4 Stein's Method in the Abstract

Comment 4.1. There are a few quite different presentations of Stein's method. One thing I'm trying to do here is to unify Stein's functional analysis approach for exchangeable pairs [Ste86] with Ross' general presentation[Ros11].

The reason I want to look at Stein's original, more abstract presentation is that I think it does a better job motivating why things work. Before I read that, the steps taken to apply Stein's method seemed like blindly doing things and it turns out they work.

Suppose we have a potentially complicated random variable X, and we believe the distribution of X is close to a "standard" distribution $\mathcal{L}_0$. Then, Stein's method allows us to compare the operators $\mathbb{E}_X$ and $\mathbb{E}_0 := \mathbb{E}_{\mathcal{L}_0}$. This is sometimes directly useful for approximating statistics of X (for example, $\mathbb{P}(X \in A) = \mathbb{E}_X \mathbf{1}_A$). However, particularly for combinatorical applications, Stein's method is most often used to bound the distance $d_{\mathcal{H}}(\mathcal{L}_X, \mathcal{L}_0)$ for some $\mathcal{H}$, where the metric $d_{\mathcal{H}}$ from Definition 2.19 is defined in terms of $\mathbb{E}_X$ and $\mathbb{E}_0$.

Stein's method is motivated by the idea of a characterizing operator.

Definition 4.1. Let $\mathcal{F}_0$ be a vector space and $\mathcal{X}_0$ be a vector space of measurable functions which contains the constant functions. We say a linear operator $T_0 : \mathcal{F}_0 \to \mathcal{X}_0$ is a *characterizing operator* for the distribution $\mathcal{L}_0$ if $\operatorname{im} T_0 = \mathcal{X}_0 \cap \ker \mathbb{E}_0$. For convenience, where there is no ambiguity we will often implicitly restrict $\mathbb{E}_0$ to $\mathcal{X}_0$, so we can write $\operatorname{im} T_0 = \ker \mathbb{E}_0$.

The following proposition shows why T_0 is called a characterizing operator.

Proposition 4.2. *If $T_0 : \mathcal{F}_0 \to \mathcal{X}_0$ is a characterizing operator and $\mathcal{X}_0$ is a determining class then $\operatorname{im} T_0 \subseteq \ker \mathbb{E}_X$ implies $\mathcal{L}_X = \mathcal{L}_0$.*

Proof. If $h \in \mathcal{X}_0$, then $h - \mathbb{E}_0 h \in \ker \mathbb{E}_0 = \operatorname{im} T_0$ so $\mathbb{E}_X[h - \mathbb{E}_0 h] = 0$. That is, $\mathbb{E}_X h = \mathbb{E}_0 h$ for all $h \in \mathcal{X}_0$, which means $\mathcal{L}_X = \mathcal{L}_0$ by the definition of a determining class. □

Proposition 4.3. *$T_0 : \mathcal{F}_0 \to \mathcal{X}_0$ is characterizing if and only if there is a linear operator $U_0 : \mathcal{X}_0 \to \mathcal{F}_0$ (called a* Stein transform*) such that the following two equations hold.*

$$\mathbb{E}_0 T_0 = 0_{\mathcal{F}_0}, \tag{4.1}$$

$$T_0 U_0 + \mathbb{E}_0 = \mathrm{id}_{\mathcal{X}_0}. \tag{4.2}$$

Proof. Suppose T_0 is a characterizing operator. Equation (4.1) is immediate. Let $\{h_i\}_{i \in \mathcal{I}}$ be a (Hamel) basis of $\mathcal{X}_0$. For each $i \in \mathcal{I}$ we have $h_i - \mathbb{E}_0 h_i \in \ker \mathbb{E}_0$ so there is some f_i (not necessarily unique) that solves $T_0 f_i = h_i - \mathbb{E}_0 h_i$. The operator U_0 can then be defined by $\sum_{i \in \mathcal{I}} a_i h_i \mapsto \sum_{i \in \mathcal{I}} a_i f_i$, satisfying (4.2).

Conversely, suppose (4.1) holds and U_0 exists satisfying (4.2). For $h \in \ker \mathbb{E}_0$ we have $T_0(U_0 h) = h$ and hence $h \in \operatorname{im} T_0$, so $\ker \mathbb{E}_0 \subseteq \operatorname{im} T_0$. Equation (4.1) immediately says that $\operatorname{im} T_0 \subseteq \ker \mathbb{E}_0$, so T_0 is a characterizing operator. □

Issue 4.2. Clean up the following remark. Maybe Banach space is too strong a requirement, can I have any topology though?

Remark 4.4. In full generality (and in accordance with [Ste86]), we do not require any topological structure on $\mathcal{F}_0$ and $\mathcal{X}_0$.

The proof of Proposition 4.3 does not ensure that the Stein transform U_0 is particularly well-behaved. In particular, if T_0 is a bounded linear operator between Banach spaces, U_0 may not be bounded. In practice, we will usually require that U_0 is well-behaved to apply Stein's

method. An equivalent condition for existence of a bounded Stein transform is that the kernel of T_0 is complementable. That is, there is a closed subspace $S \subseteq \mathcal{F}_0$ with $S \oplus \ker T_0 = \mathcal{F}_0$. For example, this condition will always be satisfied if $\mathcal{F}_0$ is (isomorphic to a) Hilbert space, by the existence of orthogonal complements. In fact, Hilbert spaces are the only spaces where every closed subspace is complementable [LT71].

Proof of claim. First assume $\ker T_0$ is complementable, with complement S. Now, $\ker \mathbb{E}_0 = \operatorname{im} T_0$ is a closed (Banach) subspace of $\mathcal{X}_0$, so T_0 restricts to a linear operator $T' : S \to \operatorname{im} T_0$ between Banach spaces. By construction, T' is bijective so by the bounded inverse theorem, T' has a bounded inverse $U' : \operatorname{im} T_0 \to S$. This gives a bounded Stein transform

$$U_0 : \mathcal{X}_0 \to \mathcal{F}_0 : h \mapsto U'(h - \mathbb{E}_0 h).$$

Conversely, if U_0 is a bounded Stein transform for T_0, let $J = \mathrm{id}_{\mathcal{F}_0} - U_0 T_0$. If $f \in \ker J \cap \ker T_0$ then $T_0 f = 0$ so $Jf = f - U_0 0 = f = 0$. That is, $\ker J \cap \ker T_0 = \{0\}$. For any $f \in \mathcal{F}_0$, we have $f = Jf + U_0 T_0 f$. Now,

$$JU_0T_0 = U_0T_0 - U_0(\mathrm{id}_{\mathcal{X}_0} - \mathbb{E}_0)T_0 = U_0T_0 - U_0T_0 + U_0\mathbb{E}_0T_0 = 0$$

and

$$T_0J = T_0 - (\mathrm{id}_{\mathcal{X}_0} - \mathbb{E}_0)T_0 = T_0 - T_0 + \mathbb{E}_0T_0 = 0,$$

so $Jf \in \ker T_0$ and $U_0 T_0 f \in \ker J$. That is, $\ker T_0 \oplus \ker J = \mathcal{F}_0$ and we conclude that $\ker T_0$ is complementable. □

We'll use Proposition 4.3 to give two important examples of characterizing operators.

Theorem 4.5. *Define $T_{\mathcal{N}}$ by $T_{\mathcal{N}} f(x) = f'(x) - x f(x)$. Let $\mathcal{X}_{\mathcal{N}} = L^1(\mathbb{R}, \mathcal{N}(0,1))$ be the set of functions $h : \mathbb{R} \to \mathbb{R}$ that satisfy $\mathbb{E}_{\mathcal{N}}|h| < \infty$ and let $\mathcal{F}_{\mathcal{N}} = T_{\mathcal{N}}^{-1}\mathcal{X}_{\mathcal{N}}$ be the set of functions $f : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}_{\mathcal{N}}|T_{\mathcal{N}} f| < \infty$. Then $T_{\mathcal{N}} : \mathcal{F}_{\mathcal{N}} \to \mathcal{X}_{\mathcal{N}}$ is a characterizing operator for $\mathcal{N}(0,1)$.*

Proof. For any $f \in \mathcal{F}_{\mathcal{N}}$, integration by parts gives

$$\mathbb{E}_{\mathcal{N}} T_{\mathcal{N}} f = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} f'(t) \, \mathrm{d}\, t - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-t^2/2} f(t) \, \mathrm{d}\, t = 0$$

so $\mathbb{E}_{\mathcal{N}} T_{\mathcal{N}} = 0$ and (4.1) holds. Then, define $U_{\mathcal{N}}$ by

$$U_{\mathcal{N}} h(x) = e^{x^2/2} \int_{-\infty}^{x} (h(t) - \mathbb{E}_{\mathcal{N}} h) e^{-t^2/2} \, \mathrm{d}\, t.$$

(obtaining this formula for $U_{\mathcal{N}}$ would involve solving a differential equation, most simply by the method of integrating factors). By the product rule and the Fundamental Theorem of Calculus, for all $h \in \mathcal{X}_{\mathcal{N}}$ we have

$$T_{\mathcal{N}} U_{\mathcal{N}} h(x) = h(x) - \mathbb{E}_{\mathcal{N}} h.$$

Hence, (4.2) holds and Proposition 4.3 completes the proof. □

Theorem 4.6. *Define* $T_{\mathrm{Po}(\lambda)}$ *by* $T_{\mathrm{Po}(\lambda)} f(k) = \lambda f(k+1) - k f(k)$. *Let* $\mathcal{X}_{\mathrm{Po}(\lambda)}$ *be the set of integer-valued functions* $h : \mathbb{N} \to \mathbb{Z}$ *that satisfy* $\mathbb{E}_{\mathrm{Po}(\lambda)} |h| < \infty$ *and let* $\mathcal{F}_{\mathrm{Po}(\lambda)} = T_{\mathrm{Po}(\lambda)}^{-1} \mathcal{X}_{\mathrm{Po}(\lambda)}$ *be the set of functions* $f : \mathbb{N} \to \mathbb{Z}$ *such that* $\mathbb{E}_{\mathrm{Po}(\lambda)} \left| T_{\mathrm{Po}(\lambda)} f \right| < \infty$. *Then* $T_{\mathrm{Po}(\lambda)} : \mathcal{F}_{\mathrm{Po}(\lambda)} \to \mathcal{X}_{\mathrm{Po}(\lambda)}$ *is a characterizing operator for* $\mathrm{Po}(\lambda)$.

Proof. For any $f \in \mathcal{F}_{\mathrm{Po}(\lambda)}$, we have

$$\mathbb{E}_{\mathrm{Po}(\lambda)} T_{\mathrm{Po}(\lambda)} f = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^{i+1}}{i!} f(i+1) - e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^i}{(i-1)!} f(i) = 0$$

so $\mathbb{E}_{\mathrm{Po}(\lambda)} T_{\mathrm{Po}(\lambda)} = 0$ and (4.1) holds. Then, define $U_{\mathrm{Po}(\lambda)}$ by

$$U_{\mathrm{Po}(\lambda)} h(k) = \frac{(k-1)!}{\lambda^k} \sum_{i=0}^{k-1} \frac{\lambda^i}{i!} \big(h(i) - \mathbb{E}_{\mathrm{Po}(\lambda)} h \big)$$

for $k \geq 1$. Substituting and simplifying gives

$$T_{\mathrm{Po}(\lambda)} U_{\mathrm{Po}(\lambda)} h(k) = h(k) - \mathbb{E}_{\mathrm{Po}(\lambda)} h,$$

so (4.2) holds and Proposition 4.3 completes the proof. □

Note that $\mathcal{H}_{\mathrm{TV}} \subseteq \mathcal{X}_{\mathcal{N}}$, where $\mathcal{H}_{\mathrm{TV}}$ is as defined in Definition 2.21. Since $\mathcal{H}_{\mathrm{TV}}$ is a determining class, $T_{\mathcal{N}}$ is a characterizing operator in the sense of Proposition 4.2. We can say the same about $T_{\mathrm{Po}(\lambda)}$ if we restrict our attention to integer-valued random variables.

Issue 4.3. Stein chose $\mathcal{X}_0 = \{h : \mathbb{E}[\mathrm{id}_{\mathbb{R}}^k |h|] < \infty \text{ for all } k\}$, for both the Poisson and normal case. I think this was to make sure T_0 and U_0 have the right domain and range. See [Ste86] page 20 and [Ste92] page 18.

The utility of the introduction of a characterizing operator is that for each $h \in \mathcal{X}_0$, Equation (4.2) allows us to to make the transformation

$$\mathbb{E}_X h = \mathbb{E}_0 h + \mathbb{E}_X T_0 U_0 h. \tag{4.3}$$

The original purpose of Stein's method was to estimate some particular $\mathbb{E}_X h$. If $\mathcal{L}_0$ was chosen to be a "simple", well-understood distribution then the term $\mathbb{E}_0 h$ should be easy to compute or estimate, and if the distribution of X was "close" to $\mathcal{L}_0$, then it should be possible to show that the remainder $\mathbb{E}_X T_0 U_0 h$ is small.

For our purposes, the main use of (4.3) is to bound $d_{\mathcal{H}}(X, \mathcal{L}_0)$ for some $\mathcal{H} \subseteq \mathcal{X}_0$. For any $\mathcal{Y} \supseteq U_0\mathcal{H}$, we have

$$d_{\mathcal{H}}(X, \mathcal{L}_0) = \sup_{h \in \mathcal{H}} |\mathbb{E}_X T_0 U_0 h| \leq \sup_{f \in \mathcal{Y}} |\mathbb{E}_X T_0 f|.$$

We have reduced the problem of bounding $d_{\mathcal{H}}(X, \mathcal{L}_0)$ to that of bounding $|\mathbb{E}_X T_0 f|$ (uniformly over $f \in \mathcal{Y}$). Especially in the cases where $\mathcal{L}_0$ is normal or Poisson and $\mathcal{H}$ is one of the standard choices in Definition 2.21, there are a number of known convenient choices of $\mathcal{Y}$, and a number of methods that are known to be effective to bound $|\mathbb{E}_X T_0 f|$.

Example 4.7. If $\mathcal{H} = \mathcal{H}_{\mathrm{TV}}$ and $\mathcal{L}_0 = \mathrm{Po}(\lambda)$, using the characterizing operator in Theorem 4.6,

then we generally choose

$$\mathcal{Y} = \Big\{ f \in \mathcal{F}_0 : \|f\|_\infty \le \min\Big\{1, \lambda^{-1/2}\Big\},\ \|\Delta f\|_\infty \le \min\big\{1, \lambda^{-1}\big\} \Big\},$$

where $\Delta f(k) = f(k+1) - f(k)$.

Issue 4.4. Maybe I can bound $\big\|U_{\mathrm{Po}(\lambda)}\big\|$ instead?

Proving that this choice of $\mathcal{Y}$ satisfies $\mathcal{Y} \supseteq U_{\mathrm{Po}(\lambda)}\mathcal{H}$ is nontrivial. But, we can prove that the constraints are of the "correct" order of magnitude.

Issue 4.5. The proof is in [BHJ92, Remark 10.2.4]. There's also a simpler proof in [BHJ92, Lemma 1.1.1] that $\|f\|_\infty \le 2\min\big\{1, \lambda^{-1/2}\big\}$ suffices. I'll revisit this later.

Proposition 4.8. *With $U_{\mathrm{Po}(\lambda)}$ as in Theorem 4.6, we have*

$$\sup_A \big\|U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A\big\|_\infty \asymp \lambda^{-1/2} \text{ as } \lambda \to \infty, \quad \sup_A \big\|U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A\big\|_\infty \asymp 1 \text{ as } \lambda \to 0,$$
$$\sup_A \big\|\Delta U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A\big\|_\infty \asymp \lambda^{-1} \text{ as } \lambda \to \infty, \quad \sup_A \big\|\Delta U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A\big\|_\infty \asymp 1 \text{ as } \lambda \to 0,$$

Proof. (Adapted from [BHJ92, Lemma 1.1.1]) For any Borel A and any $k \in \mathbb{N}$,

$$\begin{aligned}
U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A(k) &= \frac{(k-1)!}{\lambda^k}\big(\mathrm{Po}_\lambda(A \cap [k-1]) - \mathrm{Po}_\lambda(A)\,\mathrm{Po}_\lambda([k-1])\big) \\
&= \frac{(k-1)!}{\lambda^k}\bigg(\mathrm{Po}_\lambda(A \cap [k-1])\Big(1 - \mathrm{Po}_\lambda([k-1])\Big) \\
&\qquad\qquad - \Big(\mathrm{Po}_\lambda(A) - \mathrm{Po}_\lambda(A \cap [k-1])\Big)\,\mathrm{Po}_\lambda([k-1])\bigg) \\
&= \frac{(k-1)!}{\lambda^k}\Big(\mathrm{Po}_\lambda(A \cap [k-1])\,\mathrm{Po}_\lambda(\mathbb{R}\backslash[k-1]) - \mathrm{Po}_\lambda(A\backslash[k-1])\,\mathrm{Po}_\lambda([k-1])\Big).
\end{aligned}$$

Note that $\mathrm{Po}_\lambda(A \cap [k-1])$ is bounded above by $\mathrm{Po}_\lambda([k-1])$ and $\mathrm{Po}_\lambda(A\backslash[k-1])$ is bounded

above by $\mathrm{Po}_\lambda(\mathbb{R}\backslash[k-1])$, so

$$\left|U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A(k)\right| \leq \frac{(k-1)!}{\lambda^k}\,\mathrm{Po}_\lambda([k-1])\,\mathrm{Po}_\lambda(\mathbb{R}\backslash[k-1]).$$

Note that we have equality when $A = [k-1]$. If $k \asymp \lambda$ as $\lambda \to \infty$ then Stirling's approximation (unfinished...) □

Issue 4.6. I need a way to show $\mathrm{Po}_\lambda([k-1]) \asymp 1$ when k is close to λ. I imagine the typical approach would be to show that $(\mathrm{Po}(\lambda)-\lambda)/\lambda \to \mathcal{N}(0,1)$ with Lévy's continuity theorem, but that would require me to introduce more stuff in the probability revision section. Maybe I can use Stein's method itself for this! In any case I'll think about this more when I've decided on my approach to Issue 4.5.

Issue 4.7. I should have a toy example here that is amenable to several methods.

4.1 The method of exchangeable pairs

This is Stein's original approach, and is effective in wide generality for discrete random variables. In what follows, we assume X is discrete. Let Ω_X be the support of X.

Example 4.9 (adapted from [Ros11, Example 4.21])**.** We will use an example problem to illustrate the principles in this section. We will say a *fixed point* of a permutation $\sigma \in S_n$ is an index $k \in [n]$ that satisfies $\sigma(k) = k$. Let $X : S_n \to \{0\} \cup [n]$ give the number of fixed points in each permutation from S_n. We interpret X as a random variable on the underlying space $\mathcal{S}_n$.

Now, if n is large then fixed points are largely independent of each other, and each of n indices has a probability of $1/n$ to be a fixed point. So (recalling Remark 2.11), we might expect $\mathcal{L}_X$ to be "close" to $\mathrm{Po}(1)$. We will attempt to bound $d_{\mathrm{TV}}(X, \mathrm{Po}(1))$ to quantify this intuition.

The general idea is that we can use an object $\mathbf{X}$ called an exchangeable pair to construct a characterizing operator $T_{\mathbf{X}}$ for X. We then use an operator α to connect the domains of $T_{\mathbf{X}}$ and T_0 in such a way that $T_{\mathbf{X}}\alpha$ approximates T_0. We then have $\mathbb{E}_X T_0 = \mathbb{E}_X(T_0 - T_{\mathbf{X}}\alpha)$, so we can use the fact that $T_0 - T_{\mathbf{X}}\alpha$ is small to bound $\mathbb{E}_X T_0 f$.

Definition 4.10. A 2-dimensional random pair $\mathbf{X} = (X_1, X_2)$ is an *exchangeable pair* if $\mathcal{L}(X_1, X_2) = \mathcal{L}(X_2, X_1)$. We will denote the support of $\mathbf{X}$ by $\Omega_{\mathbf{X}}^{(2)}$.

That is, a pair $\mathbf{X}$ is exchangeable if exchanging the components of the pair does not change their joint distribution. In particular, the marginal distributions of X_1 and X_2 must be the same.

Comment 4.8. All presentations of Stein's method I've seen use the notation (X, X') but I think that has the potential to be confusing because the X in that pair is defined on Ω^2 whereas the original random variable X is defined on Ω.

Proposition 4.11. *There is a natural equivalence between time-homogeneous reversible Markov chains with steady-state distribution $\mathcal{L}_X$, and exchangeable pairs with margins $\mathcal{L}_X$.*

Proof. Given an exchangeable pair $\mathbf{X}$ with margins $\mathcal{L}_X$, we can define a time-homogenous Markov chain M with transition probabilities $p(x_1, x_2) = \mathbb{P}(X_2 = x_2 | X_1 = x_1)$. With $\pi(x) = \mathbb{P}(X = x)$, we then have

$$\pi(x_1)p(x_1, x_2) = \mathbb{P}(\mathbf{X} = (x_1, x_2)) = \pi(x_2)p(x_2, x_1)$$

for any $x_1, x_2 \in \Omega_X$. So, M is reversible with steady-state distribution $\mathcal{L}_X$.

Conversely, suppose we have a time-homogeneous reversible Markov chain with steady-state distribution $\mathcal{L}_X$. Let the transition probability between x and x' be $p(x, x')$ and let the probability of state x in the steady-state distribution be $\pi(x)$. We can then define an exchangeable pair $\mathbf{X}$ by

$$\mathbb{P}(\mathbf{X} = (x_1, x_2)) = \pi(x_1)p(x_1, x_2) = \pi(x_2)p(x_2, x_1) = \mathbb{P}(\mathbf{X} = (x_2, x_1))$$

and the proposition is proved. □

Definition 4.12. We say an exchangeable pair $\mathbf{X}$ is *connected* if the corresponding Markov chain is irreducible.

Remark 4.13. We are particularly interested in exchangeable pairs $\mathbf{X}$ with marginal distributions $\mathcal{L}_{X_1} = \mathcal{L}_{X_2} = \mathcal{L}_X$. If X is defined on an underlying combinatorial probability space $\big(\Omega, 2^\Omega, \mathbb{P}\big)$, it is often convenient to first construct an exchangeable pair $\mathbf{W} = (W_1, W_2)$ with margins $\mathbb{P}$, so that the vector $\mathbf{X_W} = (X(W_1), X(W_2))$ is an exchangeable pair with margins $\mathcal{L}_X$. If (W_1, W_2) is connected, then $\mathbf{X_W}$ is connected also.

Example 4.14. We continue Example 4.9. We will define a specific exchangeable pair $\mathbf{W} = (W_1, W_2)$ with margins $\mathcal{S}_n$ by

$$\mathbb{P}((W_1, W_2) = (\sigma_1, \sigma_2)) = \begin{cases} \big(n!\binom{n}{2}\big)^{-1} & \text{if } \sigma_1 = \sigma_2(i\,j) \text{ for some transposition } (i\,j) \\ 0 & \text{otherwise.} \end{cases}$$

The relation of differing by a transposition is symmetric, so $\mathbf{W}$ is indeed an exchangeable pair. The Markov chain associated with $\mathbf{W}$ has a simple interpretation. Given a random permutation σ, to make a transition in the Markov chain we just randomly choose one of the $\binom{n}{2}$ possible transpositions and compose it with σ. Because the transpositions generate S_n, the pair $\mathbf{W}$ is connected, so we can use the construction from Remark 4.13 to produce a connected exchangeable pair $\mathbf{X}$ with margins $\mathbf{W}$.

The Markov chain underlying a connected exchangeable pair can be naturally viewed as a connected one-dimensional simplicial complex. The zeroth reduced homology group $\ker \partial_0 / \operatorname{im} \partial_1$ of a connected simplicial complex has dimension zero, and this motivates the construction of a characterizing operator in a natural way. (The following theorem is self-contained and requires no knowledge of homology theory).

Theorem 4.15. *Suppose* $\mathbf{X}$ *is a connected exchangeable pair with margins* $\mathcal{L}_X$. *Let* $\mathcal{F}_X \subseteq L^1\big(\Omega_X^2, \mathcal{L}_\mathbf{X}\big)$ *be the set of functions* $f : \Omega_X^2 \to \mathbb{R}$ *which satisfy* $\mathbb{E}|f(\mathbf{X})| < \infty$, *and are antisymmetric in the sense that* $f(x_1, x_2) = -f(x_2, x_1)$. *Let* $\mathcal{X}_X = L^1(\Omega_X, \mathcal{L}_X)$ *be the set of functions* $h : \Omega_X \to \mathbb{R}$ *that satisfy* $\mathbb{E}_X|h| < \infty$.

Define $T_{\mathbf{X}} : \mathcal{F}_X \to \mathcal{X}_X$ *by* $T_{\mathbf{X}} f(x) = \sum_{x_2 \in \Omega_X} f(x, x_2) p(x, x_2) = \mathbb{E}[f(\mathbf{X}) | X_1 = x]$*, so that* $T_{\mathbf{X}} f(X)$ *is distributed as* $\mathbb{E}^{X_1} f(\mathbf{X})$*. Then* $T_{\mathbf{X}}$ *is a characterizing operator for* X*.*

Proof. To see that $\operatorname{im} T_{\mathbf{X}} \subseteq \ker \mathbb{E}_X$, fix $f \in \mathcal{F}$ and note that by the tower law of expectation (Proposition 2.15),

$$\mathbb{E}_X T_{\mathbf{X}} f = \mathbb{E}\mathbb{E}^{X_1} f(\mathbf{X}) = \mathbb{E} f(\mathbf{X}).$$

By exchangeability and antisymmetry, $\mathbb{E} f(\mathbf{X}) = \mathbb{E} f(X_2, X_1) = -\mathbb{E} f(\mathbf{X})$, so $\mathbb{E} f(X_1, X_2) = \mathbb{E}_X T_{\mathbf{X}} f = 0$. This did not require the connectedness condition. We can similarly prove that $T_{\mathbf{X}}$ is well-defined as an operator from $\mathcal{F}_X$ to $\mathcal{X}_X$: note that $\mathbb{E}_X |T_{\mathbf{X}} f| = \mathbb{E} |f(\mathbf{X})|$ so $T_{\mathbf{X}} \mathcal{F}_X \subseteq \mathcal{X}_X$.

We will next prove $\ker \mathbb{E}_X \subseteq \operatorname{im} T_{\mathbf{X}}$, but first we make some definitions. For each $x \in \Omega_X$, let h_x be the function that takes the value $\pi(x)^{-1}$ on x and is zero elsewhere, so that $h = \sum_{x \in \Omega_X} h(x) \pi(x) h_x$ for each $h \in \mathcal{X}_X$. For each $(x_1, x_2) \in \Omega_{\mathbf{X}}^{(2)}$, define $f_{x_1, x_2} \in \mathcal{F}_X$ as the function that takes the value $(\pi(x_1) p(x_1, x_2))^{-1}$ on (x_1, x_2), takes the value $-(\pi(x_2) p(x_2, x_1))^{-1}$ on (x_2, x_1), and takes the value zero elsewhere. Note that this function is antisymmetric by the reversibility of the Markov chain of $\mathbf{X}$. We have $T_{\mathbf{X}} f_{x_1, x_2} = h_{x_2} - h_{x_1}$.

Let $h \in \ker \mathbb{E}_X$, and fix an arbitrary $x^* \in \Omega_X$. By the connectedness assumption, for each $x \in \Omega_X$ there is a sequence

$$x = x^{(0)}, x^{(1)}, \ldots, x^{(k-1)}, x^{(k)} = x^*$$

with $\left(x^{(i-1)}, x^{(i)}\right) \in \Omega_{\mathbf{X}}^{(2)}$ for $i = 1, \ldots k$. Note that

$$h_{x^*} - h_x = T_{\mathbf{X}} \sum_{i=1}^{k} f_{x^{(i-1)}, x^{(i)}} =: T_{\mathbf{X}} f_x^*,$$

and it follows that

$$h = \sum_{x \in \Omega_X} h(x) \pi(x) (h_{x^*} - T_{\mathbf{X}} f_x^*) = (\mathbb{E}_X h) h_{x^*} - \sum_{x \in \Omega_X} T_{\mathbf{X}} h(x) \pi(x) f_x^*.$$

By assumption $(\mathbb{E}_X h) = 0$. If Ω_X is finite, as it will be in our applications, then it would

immediately follow that $h \in \operatorname{im} T_{\mathbf{X}}$. Otherwise we will need some functional analysis. Note that $T_{\mathbf{X}}$ is actually an isometry between a subspace $\mathcal{F}_X$ of $L^1\big(\Omega_X^2, \mathcal{L}_{\mathbf{X}}\big)$ and $L^1(\Omega_X, \mathcal{L}_X)$. First we prove that $\mathcal{F}_X$ is closed and therefore a Banach space. For any $f \in L^1\big(\Omega_X^2, \mathcal{L}_{\mathbf{X}}\big)$, let $\bar{f}$ be defined by $(x_1, x_2) \mapsto -f(x_2, x_1)$, so that $f \in \mathcal{F}_X$ implies that $f = \bar{f}$. Suppose $f_n \to f$, with $f_n \in \mathcal{F}_X$ for all $n \in \mathbb{N}$. By exchangeability we have

$$\|f_n - f\| = \mathbb{E}|f_n(\mathbf{X}) - f(\mathbf{X})| = \mathbb{E}\big|f_n(X_2, X_1) - \bar{f}(X_2, X_1)\big| = \big\|f_n - \bar{f}\big\|$$

so $f = \bar{f}$ and $f \in \mathcal{F}_X$. Finally, it is a simple fact that an isometry between Banach spaces has a closed range. For, if $(T_{\mathbf{X}} f_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\operatorname{im} T_{\mathbf{X}}$, then $(f_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathcal{F}_X$ which converges to some $f \in \mathcal{F}_X$. It follows that $T_{\mathbf{X}} f_n \to T_{\mathbf{X}} f \in \operatorname{im} T_{\mathbf{X}}$.

We have proved that $h \in \operatorname{im} T_{\mathbf{X}}$, completing the proof that $\ker \mathbb{E}_X \subseteq \operatorname{im} T_{\mathbf{X}}$. □

The final step is to choose an operator $\alpha : \mathcal{F}_0 \to \mathcal{F}_X$ in such a way that $T_{\mathbf{X}}$ can be easily compared with $T_0\alpha$.

Example 4.16. For the Poisson case in Theorem 4.6, we need to compare $\lambda f(X+1)$ with $Xf(X)$. It is often fruitful to define α by

$$\alpha f(x_1, x_2) = cf(x_2)\,\mathbf{1}\{x_2 = x_1 + 1\} - cf(x_1)\,\mathbf{1}\{x_1 = x_2 + 1\}$$

for some $c \in \mathbb{R}$. We will then have

$$T_0 f - T_{\mathbf{X}}\alpha f = f(X+1)(\lambda - c\mathbb{P}(X_2 = X_1 + 1|X_1)) - f(X_1)(X_1 - c\mathbb{P}(X_1 = X_2 + 1|X_1)).$$

Using the triangle inequality and the choice of $\mathcal{Y}$ in Example 4.7, for all $f \in \mathcal{Y}$:

$$\mathbb{E}_X T_0 f \le \min\Big(1, \lambda^{-1/2}\Big)(\mathbb{E}|\lambda - c\mathbb{P}(X_2 = X_1 + 1|X_1)| + \mathbb{E}|X_1 - c\mathbb{P}(X_1 = X_2 + 1|X_1)|).$$

This approximation is effective when

$$\mathbb{P}(X_2 = X_1 + 1|X_1) \approx \lambda/c, \tag{4.4}$$
$$\mathbb{P}(X_1 = X_2 + 1|X_1) \approx X_1/c.$$

The interpretation of these approximate equalities is that the Markov chain associated with $\mathbf{X}$ is approximately an immigration-death process. This is likely to happen when $X(\omega)$ is in some sense a statistic of the amount of local structure over the object ω, and $\mathbf{X}$ is defined by a Markov chain on Ω (as in Remark 4.13) that (uniformly) randomly disturbs local structure. The conclusion to Example 4.9 should make this clear:

Example 4.17. We continue Example 4.9, recalling the exchangeable pairs $\mathbf{W}$ and $\mathbf{X}$ from Example 4.14. The interpretation of

$$\mathbb{P}(X_1 = X_2 + 1|X_1) = \mathbb{P}(X_2 = X_1 - 1|X_1)$$

is the probability of a transposition destroying exactly one out of an existing X_1 fixed points. In order to destroy exactly one fixed point, we have to choose a fixed point to destroy, and swap it with a non-fixed-point. There are $X_1(n - X_1)$ out of $\binom{n}{2}$ transpositions that do this, so

$$\mathbb{P}(X_1 = X_2 + 1|X_1) = \frac{X_1(n - X_1)}{\binom{n}{2}}.$$

Next, we will find a formula for $\mathbb{P}(X(W_2) = X(W_1) + 1|W_1)$, noting that

$$\mathbb{P}(X_2 = X_1 + 1|X_1) = \mathbb{E}[\mathbb{P}(X(W_2) = X(W_1) + 1|W_1)|X_1].$$

In order to create exactly one fixed point, we have to choose an index k that is not fixed in W_1 (there are $n - X_1$ such) and compose W_1 with $(k\,\sigma(k))$. This creates exactly one fixed point unless $\sigma^{-1}(k) = k$, in which case it creates two. We have counted this second case twice for every transposition in the cycle decomposition of W_1. Let Y be the number of transpositions in the cycle decomposition of W_1. We have

$$\mathbb{P}(X_2 = X_1 + 1|X_1) = \frac{n - X_1 - 2\mathbb{E}[Y|X_1]}{\binom{n}{2}}.$$

In order to satisfy (4.4) as closely as possible, we choose $c = \binom{n}{2}/n$. Recalling that $\mathbb{E}X_1 = 1$, we then have

$$\begin{aligned}
\mathbb{E}|1 - c\mathbb{P}(X_2 = X_1 + 1|X_1)| &= \mathbb{E}\left[1 - \frac{n - X_1 - 2\mathbb{E}[Y|X_1]}{n}\right] \\
&= 1/n + 2\mathbb{E}Y/n, \\
\mathbb{E}|X_1 - c\mathbb{P}(X_2 = X_1 - 1|X_1)| &= \mathbb{E}\left[X_1 - \frac{X_1(n - X_1)}{n}\right] \\
&= \mathbb{E}\big[X_1^2\big]/n.
\end{aligned}$$

Now, the probability that a transposition $(i\,j)$ is in the cycle decomposition of W_1 is $(n(n-1))^{-1}$ because i must map to j out of the n possible options in $[n]$, then j must map to i out of the $n-1$ possible options in $[n]\backslash\{j\}$. There are $\binom{n}{2} = n(n-1)/2$ possible transpositions so by Remark 2.11 it follows that $\mathbb{E}Y = 1/2$.

Now, $\mathbb{E}[X_1(X_1 - 1)/2]$ is the expected number of unordered pairs of distinct fixed points in a permutation $\sigma \in \mathcal{S}_n$. For any unordered pair of distinct indices $\{i, j\}$, the probability that both are fixed is $(n(n-1))^{-1}$ because i must map to i out of the n possible options in $[n]$, then j must map to j out of the $n-1$ possible options in $[n]\backslash\{i\}$. The total number of unordered pairs is $\binom{n}{2} = n(n-1)/2$, so again applying Remark 2.11, we have $\mathbb{E}[X_1(X_1 - 1)/2] = 1/2$ and $\mathbb{E}\big[X_1^2\big] = 2$.

We conclude that $d_{\mathrm{TV}}(X, \mathrm{Po}(1)) \leq 4/n$.

Comment 4.9. The "generator method" [BC05] says that the Poisson characterizing operator can be obtained with the generator of an immigration-death process and the Normal characterizing operator can be obtained with the generator of an O-U process. Investigate the link here?

Part II

Applications

Comment 4.10. I'd like to go into a number of small examples (perhaps interspersed in the discussion of Stein's method in Part I), but I'd like to also go through a number of "big" examples. I'd like these examples to showcase

- different types of results: most applications give quantitative estimates. [Joh11] gives a non-quantitative distributional convergence result that was not previously proved using other methods. There are also results that have no connection with distribution metrics, such as the concentration inequalities in [Ros11]. In particular, the Latin rectangle example in [Ste86] is interesting in that the final result is not probabilistic.
- different types of distributions: definitely at least the Poisson and normal case, perhaps also an example of a more exotic distribution like the one in [FG12] or perturbations of Poisson/normal distributions as in [BČX07].
- different ways to apply stein's method: definitely exchangeable pairs and probably size-biasing. Maybe also Zero-bias coupling.

References

[BC05] Andrew D Barbour and Louis Hsiao Yun Chen, *An introduction to Stein's method*, vol. 4, World Scientific, 2005.

[BČX07] Andrew D Barbour, Vydas Čekanavičius, and Aihua Xia, *On Stein's method and perturbations*, arXiv preprint math/0702008 (2007).

[BHJ92] Andrew D Barbour, Lars Holst, and Svante Janson, *Poisson approximation*, Clarendon Press Oxford, 1992.

[FG12] Jason Fulman and Larry Goldstein, *Stein's method and the rank distribution of random matrices over finite fields*, arXiv preprint arXiv:1211.0504 (2012).

[Joh11] Tobias Johnson, *Exchangeable pairs, switchings, and random regular graphs*, arXiv preprint arXiv:1112.0704 (2011).

[LT71] Joram Lindenstrauss and Lior Tzafriri, *On the complemented subspaces problem*, Israel Journal of Mathematics **9** (1971), no. 2, 263–269.

[Ros11] Nathan Ross, *Fundamentals of Stein's method*, Probab. Surv **8** (2011), 210–293.

[Ste86] Charles Stein, *Approximate computation of expectations*, IMS Lecture Notes – Monograph Series, no. 7, IMS, 1986.

[Ste92] ———, *A way of using auxiliary randomization*, Probability Theory (1992), 159–180.